Retrieving Historical Manuscripts using Shape

Toni M. Rath, Victor Lavrenko and R. Manmatha*
Center for Intelligent Information Retrieval
University of Massachusetts
Amherst, MA 01002

Abstract

Convenient access to handwritten historical document collections in libraries generally requires an index, which allows one to locate individual text units (pages, sentences, lines) that are relevant to a given query (usually provided as text). Currently, extensive manual labor is used to annotate and organize such collections, because handwriting recognition approaches provide only poor results on old documents.

In this work, we present a novel retrieval approach for historical document collections, which does not require recognition. We assume that word images can be described using a vocabulary of discretized word features. From a training set of labeled word images, we extract discrete feature vectors, and estimate the joint probability distribution of features and word labels. For a given feature vector (i.e. a word image), we can then calculate conditional probabilities for all labels in the training vocabulary. Experiments show that this relevance-based language model works very well with a mean average precision of 89% for 4-word queries on a subset of George Washington's manuscripts. We also show that this approach may be extended to general shapes by using the same model and a similar feature set to retrieve general shapes in two different shape datasets.

1. Introduction

Libraries are in the transition from offering strictly paper-based material to providing electronic versions of their collections. For simple access, multimedia information, such as audio, video or images, requires an index that allows one to retrieve data, which is relevant to a given text query.

At this time, historical manuscripts like George Washington's correspondence are manually transcribed in order to provide input to a text search engine. Unfortunately, the cost of this approach is prohibitive for large collections. Automatic approaches using handwriting recognition cannot be applied (see results in [20]), since the current technology for recognizing handwriting from images has only been successful in domains with very limited lexicons and/or high redundancy, such as legal amount processing on checks and automatic mail sorting. An alternative approach called word spotting [18] which performs word image clustering is currently only computationally feasible for small collections.

Here we present an approach to retrieving handwritten historical documents from a single author, using a relevance-based language model [11, 12]. Relevance models have been successfully used for both retrieval and cross-language retrieval of text documents and more recently for image annotation[9]. In their original form, these models capture the joint statistical occurrence pattern of words in two languages, which are used to describe a certain domain (e.g. a news event).

This paradigm can be used for any signal domain, by describing images/shapes/... with *visterms* - words from a *feature vocabulary*, thus generating a "signal description language". When the joint statistical occurrence patterns of visterms and the image annotation vocabulary (e.g. word image labels) are learned, one can perform tasks such as image retrieval using text queries, or automatic annotation. While our focus here is on handwritten documents, where our signals to be retrieved are images of words, we later show that our approach can be easily adapted to work with general shapes.

In this work, we model the occurrence pattern of words in two languages using the joint probability distribution over the visterm and annotation vocabulary. From a training set of annotated images of handwritten words, we learn this joint probability distribution and perform retrieval experiments with text queries on a test set. Word images are described using a vocabulary that is derived from a set of word shape features.

Our model differs from others in a number of respects. Unlike traditional handwriting recognition paradigms [13], our approach does not require perfect recognition for good retrieval. The work presented here is also related to models used for object recognition/image annotation and retrieval

*This work was supported in part by the Center for Intelligent Information Retrieval and in part by the National Science Foundation under grant number IIS-9909073 and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Retrieving Historical Manuscripts using Shape	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Massachusetts Amherst,Department of Computer Science,140 Governors Drive,Amherst,MA,01003**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Convenient access to handwritten historical document collections in libraries generally requires an index, which allows one to locate individual text units (pages, sentences lines) that are relevant to a given query (usually provided as text). Currently, extensive manual labor is used to annotate and organize such collections, because handwriting recognition approaches provide only poor results on old documents. In this work, we present a novel retrieval approach for historical document collections, which does not require recognition. We assume that word images can be described using a vocabulary of discretized word features. From a training set of labeled word images, we extract discrete feature vectors, and estimate the joint probability distribution of features and word labels. For a given feature vector (i.e. a word image), we can then calculate conditional probabilities for all labels in the training vocabulary. Experiments show that this relevance-based language model works very well with a mean average precision of 89% for 4-word queries on a subset of George Washington's manuscripts. We also show that this approach may be extended to general shapes by using the same model and a similar feature set to retrieve general shapes in two different shape datasets.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

[6, 1, 3, 9]. However, those approaches were proposed for annotating/retrieving general-purpose photographs and primarily used color and texture as features. Here we focus on shape features for retrieval tasks, but the approach here can be extended to many shape-related retrieval and annotation tasks in computer vision.

Using this relevance-based language model, we have conducted retrieval experiments on a set of 20 pages from the George Washington collection at the Library of Congress. The mean average precision scores we achieve lie in the range from 54% to 89% for queries using 1 to 4 words (respectively). These are very good results, considering the noise in historical documents. Retrieval experiments on general shapes from the MPEG-7 and COIL-100 [16] datasets [1] yielded mean average precision of 87% and up to 97% respectively.

In the following section we discuss prior work in the field, followed by a detailed description of the relevance-based model in section 2. After briefly explaining the features used in our approach (section 3), we present line-retrieval results on the George Washington collection (section 4) and show how our retrieval approach can be extended to general shapes in section 5. Section 6 concludes the paper.

1.1. Previous Work

There are a number of approaches reported in the literature, which model the statistical co-occurrence patterns of image features and annotation words, in order to perform such diverse tasks as image annotation, object recognition and image retrieval. Mori et al. [15] estimate the likelihood of annotation terms appearing in a given image, by modeling the co-occurrence relationship between clustered feature vectors and annotation terms. Duygulu et al. [6] go one step further by actually annotating individual image regions (rather than producing sets of keywords for an image), which is in effect object class recognition. Barnard and Forsyth [1] extended Hofmann's Hierarchical Aspect Model for text and proposed a multi-modal approach to hierarchical clustering of images and words using EM. Blei and Jordan [3] extended their Latent Dirichlet Allocation (LDA) Model and proposed a Correspondence LDA model, which relates words and images.

The authors of [9] introduced the model used in this work for automatic image annotation and retrieval. With the same data and feature set, the results for image annotation were dramatically better than previous models - for example twice as good as the translation model [6]. This work extends that model to a different domain (word images in a noisy document environment), uses an improved feature representation and different attributes (shape). Shape has to be described by features that are very different from the previously utilized color and texture features. We test the model on a data set with a larger annotation vocabulary than previous experiments and a feature vector discretization that preserves more detail than the clustering algorithms which are utilized in other approaches. In addition, our application (line retrieval) uses a new retrieval model formulation. Other authors have previously suggested document-retrieval systems that do not require recognition, but queries have to be issued in the form of examples in the image domain (e.g. see [19]). To our knowledge, our system is the first to allow retrieval without recognition using text queries. We also demonstrate that this approach easily extends to more general shapes using two different data collections - the MPEG-7 and COIL-100 datasets.

All of the *image-to-word translation* approaches we are aware of, operate on image collections of good quality (e.g. the Corel image data base [6, 9]), which usually contain color and texture information. Color is known to be one of the most useful features for describing objects. Duygulu et al. [6], for example, use half of the entries in their feature vectors for color information. Images of handwritten words, on the other hand, do not generally contain color or texture information, and in the case of historical documents, the image quality is often greatly reduced.

The lack of other features makes shape a typical choice for offline handwriting recognition approaches. We make use of holistic word shape features which are justified by psychological studies of human reading[13], and are widely used in the field [5, 18, 21].

Our extension to general shape retrieval makes use of a very similar feature set and allows querying using ASCII text, which is in contrast to the many *query-by-example* retrieval approaches (see e.g. [8, 14]). The goal was not to produce the best possible shape retrieval system, but rather to demonstrate the generality of our shape retrieval model. With highly specialized shape features, such as those described in [22]), it is likely that even higher precision scores could be achieved.

2. Model Formulation

Before explaining our model in detail, we would like to provide some intuition for it. Previous research in cross-lingual information retrieval has shown that co-occurrence probabilities of words in two languages (e.g. English and Chinese) can be effectively estimated from a parallel corpus, that is, a collection of document pairs, where each document is available in two languages. Reliable estimates can be achieved even without any knowledge of the involved languages. One approach to this problem assumes that the joint distributions of, say English and Chinese words,

[1] We extracted silhouettes from the COIL-100 dataset in order to use it in our shape retrieval experiments.

are determined from a training set and may then be subsequently used to compute the probability of occurrence of the term e in an English document given the occurrence of the terms c_i in a Chinese document [23].

By analogy, word images may be described using two different vocabularies - an *image description language* - visterms - and the textual (ASCII) representation of the word. To obtain visterms, we extract features from the images and discretize them, giving us a discrete vocabulary for each word image. From a set of labeled images of words we can then estimate the joint probability $P(w, f_1 \ldots f_k)$, where w is a word label (the word "transcription") and the f_i are words from the image description language. Using the conditional density $P(w|f_1 \ldots f_k)$ we can perform retrieval of handwritten text without recognition with high accuracy.

2.1. Model Estimation

Suppose we have a collection $\mathcal{C}$ of annotated manuscripts. We will model this collection as a sequence of random variables W_i, one for each word position i in $\mathcal{C}$. Each variable W_i takes on a dual representation: $W_i = \{h_i, w_i\}$, where h_i is the image of the handwritten form at position i in the collection and w_i is the corresponding transcription of the word. As we describe in the following section, we will represent the surface form h_i as a set of discrete features $f_{i,1} \ldots f_{i,k}$ from some feature "vocabulary" $\mathcal{H}$. The transcription w_i is simply a word from the English vocabulary $\mathcal{V}$. Consequently, each random variable W_i takes values of the form $\{w_i, f_{i,1} \ldots f_{i,k}\}$. In the remaining portions of this section we will discuss how we can estimate a probability distribution over the variables W_i.

We assume that for each position i (i.e. image I_i) in the collection there exists an underlying multinomial probability distribution $P(\cdot|I_i)$ over the union of the vocabularies $\mathcal{V}$ and $\mathcal{H}$. Intuitively, our model can be thought of as an urn containing all the possible features that can appear in a representation of the word image I_i as well as all the words associated with that word image. We assume that an observed feature representation $f_1 \ldots f_k$ is the result of k random samples from this model. It follows from the urn model that the probabilities of observing $w, f_1 \ldots f_k$ are mutually independent once we pick a word image I_i with representation W_i. We further assume that actual observed values $\{w, f_1 \ldots f_k\}$ represent an i.i.d. random sample drawn from $P(\cdot|I_i)$. Then, the probability of a particular observation is given by:

$$P(W_i = w, f_1 \ldots f_k | I_i) = P(w|I_i) \prod_{j=1}^{k} P(f_j|I_i) \qquad (1)$$

Now suppose we are given an arbitrary observation $W = \{w, f_1 \ldots f_k\}$, and would like to compute the probability of that observation appearing as a random sample somewhere in our corpus $\mathcal{C}$. Because the observation is not tied to any position, we have to estimate the probability as the expectation over every position i in our entire collection $\mathcal{C}$:

$$\begin{aligned} P(w, f_1 \ldots f_k) &= E_i \left[P(W_i = w, f_1 \ldots f_k | I_i) \right] \\ &= \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} P(w|I_i) \prod_{j=1}^{k} P(f_j|I_i) \end{aligned} \qquad (2)$$

Here $|\mathcal{C}|$ denotes the aggregate number of word positions in the collection. Equation (2) gives us a powerful formalism for performing automatic annotation and retrieval over handwritten documents.

2.2. Automatic Annotation and Retrieval of Manuscripts

Suppose we are given a training collection $\mathcal{C}$ of annotated manuscripts, and a target collection $\mathcal{T}$ where no annotations are provided. Given an arbitrary handwritten image h we can automatically compute its image vocabulary (≈feature) representation $f_1 \ldots f_k$ and then use equation (2) to predict the words w which are likely to occur jointly with the features of h. These predictions would take the form of a conditional probability:

$$P(w|f_1 \ldots f_k) = \frac{P(w, f_1 \ldots f_k)}{\sum_{v \in \mathcal{V}} P(v, f_1 \ldots f_k)} \qquad (3)$$

This probability could be used directly to annotate new handwritten images with highly probable words. We provide a brief evaluation for this kind of annotation in section 4.2. However, if we are interested in retrieving sections of manuscripts we can make another use of equation (3).

Suppose we are given a user query $Q = q_1 \ldots q_m$. We would like to retrieve sections $S \subset \mathcal{T}$ of the target collection that contain the query words. More generally, we would like to *rank* the sections S by the probability that they are relevant to Q. One of the most effective methods for ranked retrieval is based on the statistical language modeling framework [17]. In this framework, sections S of text are ranked by the probability that the query Q would be observed during i.i.d. random sampling of words from S:

$$P(Q|S) = \prod_{j=1}^{m} \hat{P}(q_j|S) \qquad (4)$$

In text retrieval, estimating the probability $\hat{P}(q_j|S)$ is straightforward – we just count how many times the word q_j actually occurred in S, and then normalize and smooth

the counts. When we are dealing with handwritten documents we do not know what words did or did not occur in a given section of text. However, we can use the conditional estimate provided by equation (3):

$$\hat{P}(q_j|S) = \frac{1}{|S|}\sum_{o=1}^{|S|} P(q_j|f_{o,1}\ldots f_{o,k}) \quad (5)$$

Here $|S|$ refers to the number of word-images in S, the index o goes over all positions in S, and $f_{o,1}\ldots f_{o,k}$ represent a set of features derived from the word image in position o. Combining equations (4) and (5) provides us with a complete system for handwriting retrieval.

2.3. Estimation Details

In this section we provide the estimation details necessary for a successful implementation of our model. In order to use equation (2) we need estimates for the multinomial models $P(\cdot|I_i)$ that underly every position i in the training collection $\mathcal{C}$. We estimate these probabilities via smoothed relative frequencies:

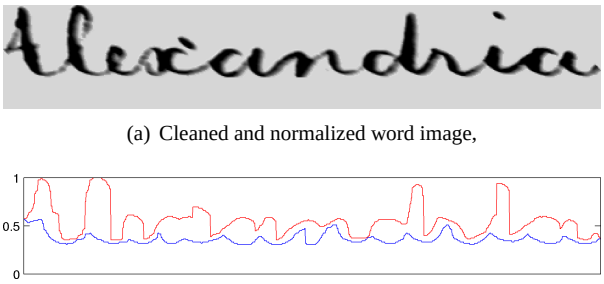
$$\begin{aligned}\hat{P}(x|I_i) = &\frac{\lambda}{1+k}\delta(x \in \{w_i, f_{i,1}\ldots f_{i,k}\}) \\ &+ \frac{(1-\lambda)}{(1+k)|\mathcal{C}|}\sum_{l\in\mathcal{C}}\delta(x \in \{w_l, f_{l,1}\ldots f_{l,k}\}) \quad (6)\end{aligned}$$

where $\delta(x \in \{w, f_1\ldots f_k\})$ is a set membership function, equal to one if and only if x is either w or one of the feature vocabulary terms $f_1\ldots f_k$. Parameter λ controls the degree of smoothing on the frequency estimate and can be tuned empirically.

3. Features and Discretization

The word shape features we use in this work are described in [10] *(the feature section of that article was submitted to the conference review system as an anonymized supplemental file)*. They are holistic word shape features, ranging from word image width/height to low-order discrete Fourier transform (DFT) coefficients of word shape profiles (see Figure 1). This feature set allows us to represent each image of a handwritten word with a continuous-space feature vector of constant length.

With these feature sets we get a 26-dimensional vector for word shapes. These representations are in continuous-space, but the relevance model requires us to represent all feature vectors in terms of a *visterm* vocabulary of fixed size. Previous approaches [9] use clustering of feature vectors, where each cluster corresponds to one visterm. However, this approach is rather aggressive, since it considers words or shapes to be equal if they fall into the same cluster.



(a) Cleaned and normalized word image,

(b) resulting upper and lower profile features displayed together.

Figure 1: Two of the three shape profile features.

We chose a discretization method that preserves a greater level of detail, by separately binning each dimension of a feature vector. Whenever a feature value falls into a particular bin, an associated visterm is added to the discrete-space representation of the word or shape. We used two overlapping binning schemes - the first divides each feature dimension into 10 bins while the second creates an additional 9 bins shifted by half a bin size. The additional bins are used to assign similar feature values to at least one same visterm. After discretization, we have 52 visterms per word image. The entire visterm vocabulary contains $26 \cdot 19 = 494$ entries.

4. Handwriting Retrieval Experiments

We will discuss two types of evaluation. First, we briefly look at the predictive capability of the annotation as outlined in section 2. We train a model on a small set of annotated manuscripts and evaluate how well the model was able to annotate each word in a held-out portion of the dataset. Then we turn to evaluating the model in the context of ranked retrieval.

The data set we used in training and evaluating our approach consists of 20 manually annotated pages from George Washington's handwritten letters. Segmenting this collection yielded a total of 4773 images, from which the majority contain exactly one word. An estimated 5-10% of the images contain segmentation errors of varying degrees: parts of words that have faded tend to get missed by the segmentation, and occasionally images contain 2 or more words or only a word fragment.

4.1. Evaluation Methodology

Our dataset comprises 4773 total word occurrences arranged on 657 lines. Because of the relatively small size of the dataset, all of our experiments use a 10-fold randomized cross-validation, where each time the data is split into a 90% training and 10% testing sets. Splitting was performed on a line level, since we chose lines to be our retrieval unit.

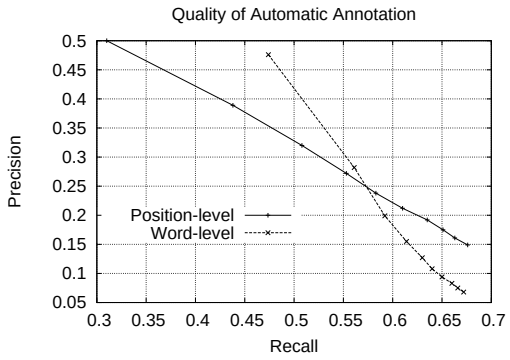


Figure 2: Performance on annotating word images with words.

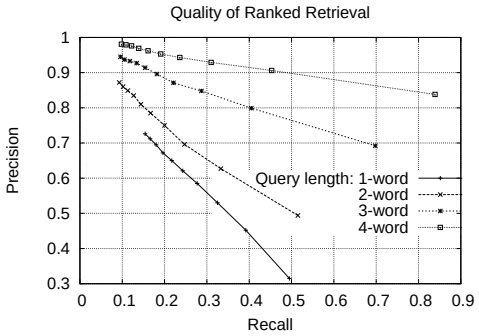


Figure 3: Performance on ranked retrieval with different query sizes.

Prior to any experiments, the manual annotations were reduced to the root form using the Krovetz morphological analyzer. This is a standard practice in information retrieval, it allows one to search for semantically similar variants of the same word. For our annotation experiments we use every word of the 4773-word vocabulary that occurs in both the training and the testing set. For retrieval experiments, we remove all function words, such as "of", "the", "and", etc. Furthermore, to simulate real queries users might pose to our system, we tested all possible combinations of 2, 3 and 4 words that occurred on the same line in the testing, but not necessarily in the training set. Function words were excluded from all of these combinations.

We use the standard evaluation methodology of information retrieval. In response to a given query, our model produces a ranking of all lines in the testing set. Out of these lines we consider only the ones that contain all query words to be relevant. The remaining lines are assumed to be non-relevant. Then for each line in the ranked list we compute *recall* and *precision*. Recall is defined as the number of relevant lines above (and including) the current line, divided by the total number of relevant lines for the current query. Similarly, precision is defined as number of above relevant lines divided by the rank of the current line. Recall is a measure of what percent of relevant lines we found, and precision suggests how many non-relevant lines we had to look at to achieve that recall. In our evaluation we use plots of precision vs. recall, averaged over all queries and all cross-validation repeats. We also report Mean Average Precision, which is an average of precision values at all recall points.

4.2. Discussion of Results

Figure 2 shows the performance of our model on the task of assigning word labels to handwritten images. We carried out two types of evaluation. In **position-level** evaluation, we generated a probability distribution $P(w|f_{i,1}\ldots f_{i,k})$ for every position i in the testing set. Then we looked for the rank of the correct word w in that distribution and averaged the resulting recall and precision over all positions. Since we did not exclude function words at this stage, position-level evaluation is strongly biased toward very common words such as "of", "the" etc. These words are generally not very interesting, so we carried out a **word-level** evaluation. Here for a given word w we look at the ranked list of all the positions i in the testing set, sorted in the decreasing order of $P(w|f_{i,1}\ldots f_{i,k})$. This is similar to running w as a query and retrieving all *positions* in which it could possibly occur. Recall and precision were calculated as discussed in the previous section.

From the graphs in Figure 2 we observe that our model performs quite well in annotation. For position-level annotation, we achieve 50% precision at rank 1, which means that for a given position i, half the time the word w with the highest conditional probability $P(w|f_{i,1}\ldots f_{i,k})$ is the correct one. Word-oriented evaluation also has close to 50% precision at rank 1, meaning that for a given word w the highest-ranked position i contains that word almost half the time. Mean Average Precision values are 54% and 52% for position-oriented and word-oriented evaluations respectively.

Now we turn our attention to using our model for the task of retrieving relevant portions of manuscripts. As discussed before, we created four sets of queries: 1, 2, 3 and 4 words in length, and tested them on retrieving line segments. Our experiments involve a total of 1950 single-word queries, 1939 word pairs, 1870 3-word and 1558 4-word queries over 657 lines. Figure 3 shows the recall-precision graphs. It is very encouraging to see that our model performs extremely well in this evaluation, reaching over 90%

mean precision at rank 1. This is an exceptionally good result, showing that our model is nearly flawless when even such short queries are used. Mean average precision values were 54%, 63%, 78% and 89% for 1-, 2-, 3- and 4-word queries respectively. Figures 4 and 5 show two retrieval results with variable-length queries. We have implemented a demo web-interface for our retrieval system, which can be found at <*URL omitted for review process*>.

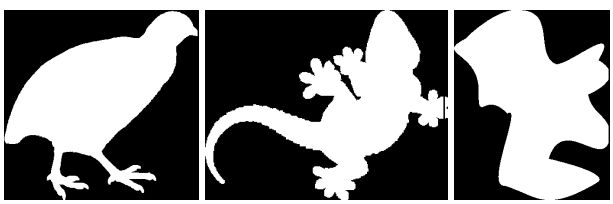
5. General Shapes

We performed proabilistic annotation and retrieval experiments on the MPEG-7 shape and COIL-100 datasets to demonstrate the extensibility of our model and features to general shapes.

The feature set for the retrieval of general shapes was adapted by removing the estimate for the number of descenders (word-specific) and the image height and width features (redundant after shape normalization). In order to get more accurate representations of the shapes, the projection profile and upper/lower profiles were complemented by also calculating them for the shape at a 90 degree rotation angle. With these feature sets we get a 44-dimensional vector for general shapes as compared to the 26-dimensional vector for word shapes. Discretization as before gives 88 visterms per shape with a total visterm vocabulary size of $44 \cdot 19 = 861$.

5.1. MPEG-7 Dataset

The MPEG-7 dataset (see Figure 6) consists of 1400 shape images of 70 shape categories, with 20 examples per category (e.g. "apple"). To prepare the shapes for the feature extraction, we performed a closing operation on each shape, rotated it so that its principal axis is oriented horizontally and normalized its size. After the feature vectors were extracted and discretized into visterms (see section 3), we performed retrieval experiments using 10-fold cross-validation. For the retrieval experiments, we ran 70 ASCII queries on the testing set. Each of the unique 70 shape category labels serves as a query.



(a) bird, (b) lizzard, (c) Misk.

Figure 6: MPEG7 shape examples with annotations (from file names).

For each cross-validation run we have a 90% training/10% testing split of the entire dataset. We performed retrieval experiments on the training portion in order to determine the smoothing parameters λ for the visterm and annotation vocabularies. The smoothing parameters that yieldeded the best retrieval performance are then used for retrieval on the testing split.

Mean average precision	Standard deviation
87.24%	4.24%

Table 1: Mean average precision results for the retrieval experiments on the MPEG-7 shape dataset averaged over 10 cross-validation runs with standard deviation.

Table 1 shows the mean average precision results we achieved with the 10 cross-validation runs. Even with this very simple extension of our word-features and the same model we can get very high retrieval performance at 87% mean average precision. It is important to note that in contrast to the common query-by-content retrieval systems, which require some sort of shape drawing as a query, we have actually learned each shape category concept, and can retrieve similar shapes with an ASCII query.

5.2. COIL-100 Dataset

In the MPEG-7 dataset, each shape is usually seen from the side. For increased complexity we turned to the COIL-100 dataset [16]. This dataset contains 7200 color images of 100 household objects and toys. Each object was placed on a turntable and an image was taken for every 5 degrees of rotation, resulting in 72 images per object. We converted the color images into shapes by binarizing the images (see Figure 7 for examples) and normalizing their sizes. In order to facilitate retrieval using text queries, each object was labeled with one of 45 class labels (these are also used as queries).

After extracting features and turning them into visterms, we performed retrieval experiments with varying numbers of training examples per object category. The number of examples per object are (evenly spaced througout 360 degrees of rotation): 1, 2, 4, 8, 18, and 36. Once the training examples are selected, we pick 9 shapes per object at random from the remaining shapes. This set, which contains a total of $9 \cdot 100 = 900$ shapes, is used to train the smoothing parameters of the retrieval model. From the remaining shapes, another 9 shapes per object are selected at random to form the testing set on which we determine the retrieval performance.

Figure 8 shows the mean average precision results obtained in this experiment ("all queries" plot). Unfortunately we were not able to show any retrieval examples due to space constraints. The "reduced query set" plot shows the

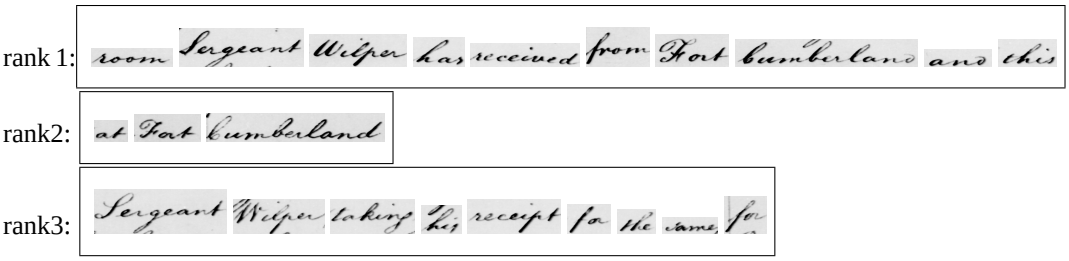


Figure 4: Retrieval result for the 4-word query “sergeant wilper fort cumberland” (one relevant line in collection).

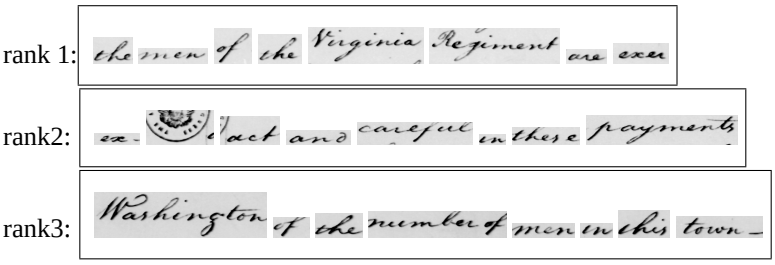
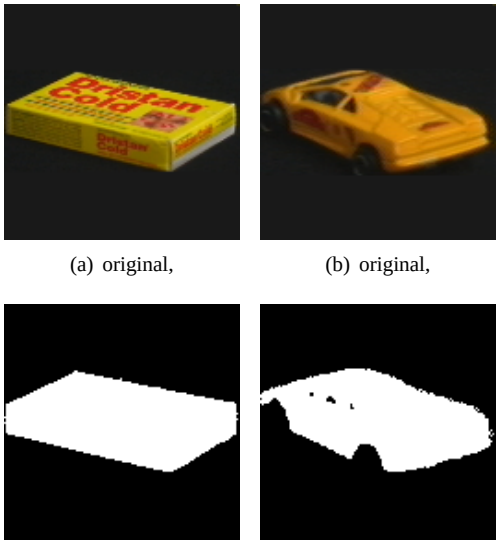


Figure 5: Retrieval result for the 3-word query “men virginia regiment” (one relevant line in collection).



(a) original, (b) original,

(c) extracted shape “box”, (d) extracted shape “car”.

Figure 7: COIL-100 dataset examples: original color images and extracted shapes with our annotations.

same experiment, where queries are omitted for objects that are invariant under the turntable rotation performed during the COIL-100 dataset acquisition. As expected, the average precision scores are slightly lower, but the differences become negligible when there are many examples per object (for 36 examples, the “reduced query set” is actually about .5% better than “all queries”).

These results are very encouraging, since they indicate we can perform satisfactory retrieval at around 80% mean average precision (m.a.p.) for 8 examples per object (45 degrees apart) and high performance retrieval at 97% m.a.p. for 36 examples per object (10 degrees apart). Note that this is done exclusively on shape images (without using any intensity information). Clearly, if other information and a more specialized feature set were used, even higher precision scores could be achieved.

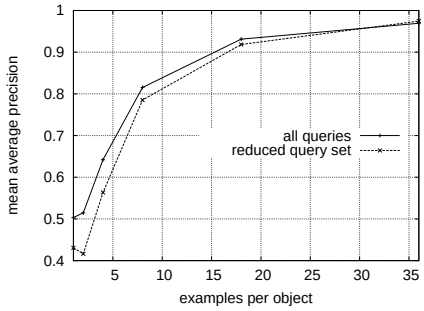


Figure 8: Retrieval results on the COIL-100 dataset for different numbers of examples per object. The reduced query set excludes queries for objects that appear invariant under the rotation performed during the dataset acquisition.

6. Summary and Conclusion

We have presented a relevance-based language model for the retrieval of handwritten documents and general shapes.

Our model estimates the joint probability of occurrence of annotation and feature vocabulary terms in order to perform probabilistic annotation and retrieval of handwritten words (documents) and general shapes. Our approach is the first to use shape-based features, and we presented appropriate shape representation, discretization and retrieval techniques. The results for the retrieval of lines of handwritten text indicate performance at a level that is practical for real-world applications.

Future work will include a retrieval system for a larger collection with page retrieval. Extending the collection could require more features in order to discriminate better between similar words. Lastly, we also plan to work on improved retrieval models.

Acknowledgments

We would like to thank the Library of Congress for providing the scanned images of the George Washington collection.

References

[1] K. Barnard and D. Forsyth: *Learning the Semantics of Words and Pictures*. In: Proc. of the Int'l Conf. on Computer Vision, vol. 2, Vancouver, Canada, July 9-12, 2001, pp. 408-415.

[2] A. Berger, and J. Lafferty: *Information Retrieval as Statistical Translation*. In: Proc. of the 22nd Annual Int'l SIGIR Conf., 1999, pp. 222-229.

[3] D. M. Blei, and M. I. Jordan: *Modeling Annotated Data*. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf., Toronto, Canada, July 28-August 1, 2003, pp. 127-134.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan: *Latent Dirichlet Allocation*. Journal of Machine Learning Research **3** (2003) 993-1022.

[5] C.-H.Chen: *Lexicon-Driven Word Recognition*. In: Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition 1995, Montréal, Canada, August 14-16, 1995, pp. 919-922.

[6] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth: *Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary*. In: Proc. of the 7th European Conf. on Computer Vision, Copenhagen, Denmark, May 27-June 2, 2002, vol. 4, pp. 97-112.

[7] D. Hiemstra: *Using Language Models for Information Retrieval*. Ph.D. dissertation, University of Twente, Enschede, The Netherlands, 2001.

[8] A. K. Jain and A. Vailaya: *Shape-Based Retrieval: A Case Study With Trademark Image Databases*. Pattern Recognition **31**:9 (1998) 1369-1390.

[9] J. Jeon, V. Lavrenko and R. Manmatha: *Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models*. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf., Toronto, Canada, July 28-August 1, 2003, pp. 119-126.

[10] V. Lavrenko, T. M. Rath and R. Manmatha: *Holistic Word Recognition for Handwritten Historical Documents*. In: Proc. of the Int'l Workshop on Document Image Analysis for Libraries (DIAL), Palo Alto, CA, January 23-24, 2004 (to appear).

[11] V. Lavrenko and W. B. Croft: *Relevance-Based Language Models*. In: Proc. of the 24th Annual Int'l SIGIR Conf., New Orleans, LA, September 9-13, 2001, pp. 120-127.

[12] V. Lavrenko, M. Choquette and W. B. Croft: *Cross-Lingual Relevance Models*. In: Proc. of the 25th Annual Int'l SIGIR Conf., Tampere, Finland, August 11-15, 2002, pp. 175-182.

[13] S. Madhvanath and V. Govindaraju: *The Role of Holistic Paradigms in Handwritten Word Recognition*. Trans. on Pattern Analysis and Machine Intelligence **23**:2 (2001) 149-164.

[14] G. Mori, S. Belongie and J. Malik: *Shape Contexts Enable Efficient Retrieval of Similar Shapes*. In: Proc. of the Conf. on Computer Vision and Pattern Recognition vol. 1, Kauai, HI, December 9-14, 2001, pp. 723-730.

[15] Y. Mori, H. Takahashi and R. Oka: *Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words*. In: 1st Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM), Orlando, FL, October 30, 1999.

[16] S. A. Nene, S. K. Nayar and H. Murase: *Columbia Object Image Library (COIL-100)*. Technical Report CUCS-006-96, February 1996.

[17] J.M. Ponte and W.B. Croft: *A Language Modeling Approach to Information Retrieval*. In: Proc. of the 21st Annual Int'l SIGIR Conf., Melbourne, Australia, August 24-28, 1998, pp. 275-281.

[18] T. M. Rath, R. Manmatha: *Word Image Matching Using Dynamic Time Warping*. In: Proc. of the Conf. on Computer Vision and Pattern Recognition, Madison, WI, June 18-20, 2003, vol. 2, pp. 521-527.

[19] C. L. Tan, W. Huang and Y. Xu: *Imaged Document Text Retrieval without OCR*. Trans. on Pattern Analysis and Machine Intelligence **24**:6 (2002) 838-844.

[20] C. I. Tomai, B. Zhang and V. Govindaraju: *Transcript Mapping for Historic Handwritten Document Images*. In: Proc. of the 8th Int'l Workshop on Frontiers in Handwriting Recognition 2002, Niagara-on-the-Lake, ON, August 6-8, 2002, pp. 413-418.

[21] Ø. D. Trier, A. K. Jain and T. Taxt: *Feature Extraction Methods for Character Recognition - A Survey*. Pattern Recognition **29**:4 (1996) 641-662.

[22] R. C. Veltkamp and M. Hagedoorn: *State-of-the-Art in Shape Matching*. Technical Report UU-CS-1999-27, Utrecht University, the Netherlands, 1999.

[23] J. Xu, R. Weischedel and C. Nguyen: *Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval*. In: Proc. of the 24th Annual Int'l ACM-SIGIR Conf. on Research and Development in Information Retrieval, New Orleans, LA, September 9-13, 2001, pp. 105-110.

[24] C. Zhai: *Risk Minimization and Language Modeling in Text Retrieval*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 2002.